View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DCU Online Research Access Service

Link Anchors in Images: Is There Truth?

Robin Aly[1], Kevin McGuinness[2], Martijn Kleppe[3], Roeland Ordelman[1]
Noel O'Conner[2], Franciska de Jong[1]
[1]: University of Twente, [2]: Dublin City University, [3]: Erasmus University Rotterdam

ABSTRACT

While automatic linking in text collections is well understood, little is known about links in images. In this work, we investigate two aspects of anchors, the origin of a link, in images: 1) the requirements of users for such anchors, e.g. the things users would like more information on, and 2) possible evaluation methods assessing anchor selection algorithms. To investigate these aspects, we perform a study with 102 users. We find that 59% of the required anchors are image segments, as opposed to the whole image, and most users require information on displayed persons. The agreement of users on the required anchors is too low (often below 30%) for a ground truth-based evaluation, which is the standard IR evaluation method. As an alternative, we propose a novel evaluation method based on improved search performance and user experience.

1. INTRODUCTION

Links help to satisfy spontaneous information needs of users that occur when they inspect the content of documents [2]. Although the importance of links is widely acknowledged, manually creating such links is time consuming, particularly for large document collections. This has motivated researchers to investigate methods for automatic link generation [5]. Developing link generation algorithms necessitates a solid understanding of the requirements that users have for links and evaluation methods, which has been investigated for links in text collections [9]. However, little is known about the requirements of automatically generated links originating from images and their evaluation. The main contribution of this paper is to narrow this knowledge gap.

A link consists of an anchor, the area of an image for which users need information, and link destination(s) where this need is satisfied. Given an image, a link generation algorithm first selects potentially interesting anchors, and determines their link destinations afterwards. This paper focuses on properties of anchors, and possible evaluation methods for automatic selection algorithms.

First, it is important to know what kind of anchors users want. Traditionally, only whole images are assumed possible anchors [2]. However, the expression "a picture is worth a thousand words" suggests that anchors can also be image segments. Therefore, we investigate to what extent users are interested in segments compared to whole image. Furthermore, anchor selection algorithms have to employ multimedia extraction techniques, which are often content type specific. For example, techniques to detect people and events differ largely. Therefore, we also investigate what types of information needs users have in images.

Given an anchor selection algorithm, we need an evaluation method for its quality. The standard evaluation method in IR is to compare an algorithm's output with a previously established ground truth (e.g. relevance judgments), which makes experiments repeatable at minimal evaluation costs. This evaluation method has also been adapted to linking in text collections [3]. However, ground truth-based evaluation relies on a basic assumption: there is a universal truth on which a large user group agrees. Low agreement on anchors inside images indicates that an alternative evaluation method is needed. We investigate the agreement of users on the anchors they need in images, and find that agreement is too low to use a ground truth-based evaluation method.

Link structures have also been evaluated based on users' experience [7, 6]. These evaluation methods are however limited to small collections and expensive to conduct, because they require users freely browsing the collection. We propose a new evaluation method that combines ground truth-based and experience-based evaluation methods: the quality of a link structure is assessed by the improvement on interactive search effectiveness (measured by relevance judgments) and users' experience when interacting with the system.

We approach the above research questions by a large-scale user-study of 102 participants using a self-created image collection of roughly 895 images where participants were asked to select anchors in images.[1] Given the early stage of research in automatic link generation in images, this work clearly has to limit its scope. We do not investigate the evaluation of link destinations, although they are clearly important. We found in initial studies that it is difficult for users to determine link destinations, which therefore requires another investigation approach. Furthermore, although user studies ideally cover multiple collections, we believe that our collection contains sufficient variation to allow general

[1]The collection together with the results of the user study will be made publicly available after the review process.

Environment	Immigration / Society	Politicians
Fukushima	Refugees	Sarkozy
Tsunami Japan	"Asylum Seeker"	Merkel
Nuclear Japan	Foreigner	"David Cameron"
Earthquake Japan	Immigrant	"Nicolas Sarkozy"
Katrina	"Multicultural Society"	"Angela Merkel"
"New Orleans" hurricane	"Lampedusa Refugee"	Cameron
"New Orleans" Superdome	"Italy refugee"	"European summit"

Table 1: The seed queries that we used to generate the data set of images, which users had to annotate.

statements. Finally, due to the lack of automatic extraction tools, we are only able to outline the described evaluation method without validating it.

The remainder of this paper is organized as follows. Section 2 discusses the connections of this paper to related work. Section 3 describes the design of the user-survey. Section 4 describes the results and discusses them. Section 5 proposes an alternative way of assessing link structures. Section 6 draws conclusions from those findings.

2. RELATED WORK

In the following, we describe related work to this paper. Fountain et al. [2] was one of the first to use links as a generic connection between two documents of arbitrary modality. Subsequent work investigated the requirements users have for such links, see for example [9] on the requirements of links in text. Other work proposes methods to automatically select anchors (sequences of words) in text documents [5]. Simple methods for anchor text selection, like the inverse document frequency of terms, fit the user needs well. So far, the requirements of links in images have received far less attention, and no automatic anchor selection algorithms have been proposed. This paper tries to narrow this knowledge gap.

Ground truth-based evaluation is the de-facto standard in information retrieval where the truth consists of relevance assessments. Voorhees [8] investigates the influence of agreement among relevance assessors on ad-hoc search performance. She finds that a relatively low agreement (Jaccard index of roughly 0.60), does not affect performance measures significantly. However, Al-Maskari et al. [1] finds that similar agreement changes the search performance in other search tasks. In this paper, we find distinctly lower agreement among users selecting anchors in images, which prohibits ground truth-based evaluation.

Link structures are also evaluated based on user experience [7, 6]. Here, users are asked to browse a collection without giving them a particular task. The users experience is then evaluated through questionnaires. There are two disadvantages of such evaluation methods: they are expensive and user dependent, because users have to browse the collection for a long time without intrinsic motivation. Our proposed method evaluates link structures through users performing a search task, therefore reducing costs and increasing user focus.

3. USER STUDY DESIGN

In the following, we describe the user study that we used to obtain our results. Choosing an appropriate collection for a user study is important since it can have a significant effect on the information needs of the users. The collection should contain a high number of images with interesting content for many users. We investigated several well-known image collections for this survey, but found that existing collections either did not contain a sufficient number of interesting images, or were encumbered by copyright restrictions that made them unsuitable for an online user study. We therefore decided to compile a new collection for this study using images from Flickr.

We asked a humanity scholar with expertise in the effect of images on humans to formulate a set of 21 queries based on the topics *the environment*, *immigration and society*, and *politicians*, to aid us in selecting interesting images. The scholar first investigated the results returned from Flickr before adding a query to the final query set, which is shown in Table 1. Afterwards, we downloaded the first five result pages for each of these queries from Flickr, which resulted in a collection of 895 images.

Furthermore, the design of the user interface is important. Figure 1 (a) shows the main page of the user study which displays one image at a time. The user had three options: 1) to select the whole image, 2) to make a rectangular selection, and finally 3) to skip to the next image, if no more information needs occurred. The users were instructed to state as many information needs as they found suitable. If the user selected the whole image or a rectangle, the dialog in Figure 1 (b) was shown. Here, users were asked to specify the information needs they had concerning this selection. To be later able to group information needs, we asked the users to classify them in one of the following types, motivated by existing content-extraction techniques: 1) Persons, 2) Objects, 3) Places, 4) Events, and 5) Other. Each user had to process ten images, and we ensured that each image was at least processed by five users. As a post-processing step, we manually grouped the selections with a high overlap that we thought represented the same content and only differed by the user's precision in selecting the interesting areas.

4. RESULTS AND DISCUSSION

The user study was carried out on-line, and was advertised on social networking sites, among colleagues and friends, and in classes taught by the authors. The survey was completed by 102 users from eight countries with three main types of professions: students, researchers, and media professionals.

Figure 2 (a) shows a histogram of the number of selections in each image. The figure shows that in roughly 150 cases the user did not have any information need (anchor) for an image shown. This may be due to low interest in the image content, or that the user simply did not need additional information. Out of all selections, 59% concerned an image segment (as

(a) (b)

Figure 1: Link anchor selection and formulation of information needs.

opposed to selections of the whole image). Figure 2 (b) shows the distribution of the types of information needs the users had for these images. We can see that users were distinctly interested in people occurring in images.

Following [8], we investigated the agreement among users on anchors by the Jaccard index using overlap in the selected anchors:

$$Agreement_{i,j} = \frac{|Selections_i \cap Selections_j|}{|Selections_i \cup Selections_j|} \quad (1)$$

where $Selections_i$ and $Selections_j$ are the selected anchors of user i and j respectively. Because a user could select multiple anchors in an image, the average agreement for an image was then calculated as the average of agreements over all pairs of users that viewed the image:

$$AvgAgreement = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Agreement_{i,j} \quad (2)$$

where n is the number of users assigned to the image. Figure 3 shows a histogram of the average user agreement. Of all images, 75% had an average agreement of below 31%. We created similar histograms for the agreement specific to the five previously mentioned information need types and professions, which did not show large differences to the distribution in Figure 3.

We now discuss the results of the study. The users of the study were interested in the collection, which can be seen from the number of anchors, and they were mainly interested in persons. Other information types are, roughly speaking, equally important.

The results on the agreement indicate that users often have only a single information need per image and often do not agree on the anchors in an image. Given such low agreement, a ground truth-based evaluation of selected anchors in images will only reflect a certain user group and will not generalize to a larger user group, as will be any evaluation measure based upon this truth.

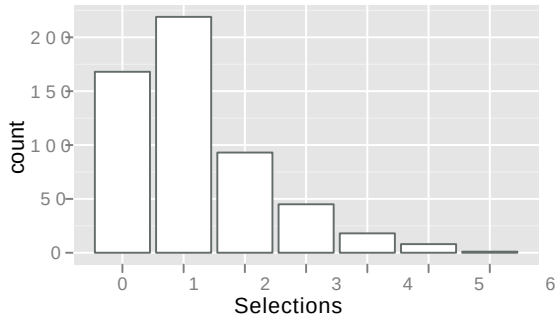
5. SEARCH-BASED EVALUATION / FUTURE WORK

Given the low agreement of users about anchors in images, we propose that anchor selections cannot be evaluated using a ground truth. Instead, we sketch an alternative evaluation method for link structures (link anchors destinations) based on their usefulness for search and users' experience, which we plan to investigate in the future. Similarly to the method by McDonald and Stevenson [4] for links in text collections, we assume that users' browsing experience is similar to their experience when using links for searching. Users are asked to perform an interactive search task where they are supposed to find relevant images. According to the latin square method, we iteratively present the user a system, which uses links, and the same system, but without links. We can then measure the gain in search effectiveness by the difference of the two search performance measures. Furthermore, we store the links followed during the interactive search session in a log file. Similar to the evaluation of links in text collections [7, 6], we then use the number of followed links compared to the overall available links in a visited document as a measure for the user experience of the underlying link structure. We suggest that the combination of search effectiveness and user experience will accurately represent the quality of a link structure.

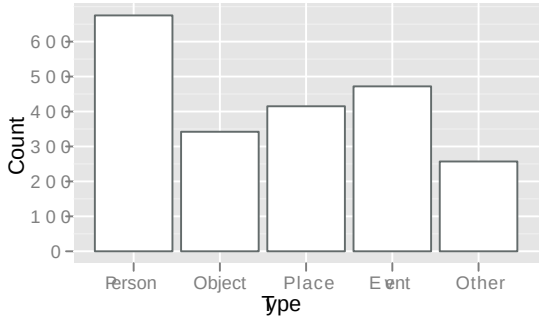
6. CONCLUSIONS

This paper investigated user requirements for links in (parts of) images and evaluation methods of their automatic selection. For the requirements, we were mainly interested in whether users want to link whole images or parts of them, and the types of information needs they have when looking at images, which should be satisfied by links. For the evaluation method, we investigated whether the agreement among users was high enough to use a ground truth-based evaluation style.

We performed a user study where participants were asked to select rectangular segments of images or the whole image to which they would like to additional information to be linked. We created a new collection of images with the help of a humanities scientist to ensure the images were sufficiently interesting. We found that 59% of the users were interested in image segments. Furthermore, they were mainly interested in information on persons. The agreement among users was low: 75% of all images had an average agreement of below 31%, measured by the Jaccard index, on the required anchors. There was no significant difference when



(a)



(b)

Figure 2: Histograms showing (a) the number of selections in per image and (b) information need types per selection (note that users were able to make multiple selections each having multiple information needs per image).

only considering certain information types (Persons, Objects, etc.). This provides evidence that ground truth-based evaluation is not suitable for the assessment of automatically anchor selection algorithms.

As an alternative, we proposed a method to evaluate link structures based on a mixture of change in search effectiveness and user experience. Users are involved in an interactive search task using a system, which employs links and the same system without links. The difference in search performance is therefore a measure for the increased search effectiveness of links. Additionally, by recording the links used during search, existing user-experience measures for link structures can be calculated. The quality of a link structure is then represented by the mixture of both measures.

We believe that the presented study is an important step towards automatically selecting link anchors in images. However, user behavior is collection dependent, we plan to investigate the agreement for other collections in the future. Furthermore, the results of the survey can also help to determine *why* certain user groups, e.g. depending on age or nationality, require certain links. Potentially, this could guide the development of anchor selection methods in the future.

Acknowledgements

This work was co-funded by the EU FP7 Project AXES ICT-269980. We want to thank the anonymous reviewers for their valuable input.

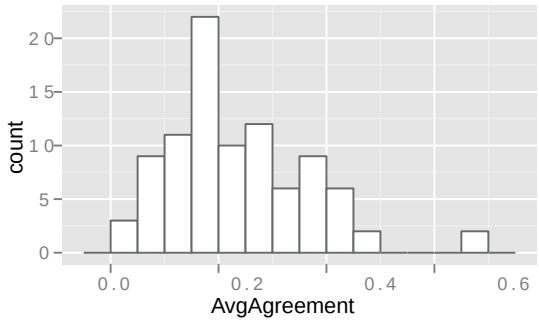


Figure 3: Histogram of the average agreement among users.

References

[1] A. Al-Maskari, M. Sanderson, and P. Clough. Relevance judgments between trec and non-trec assessors. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 683–684, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390450.

[2] A. Fountain, W. Hall, I. Heath, and H. Davis. Microcosm: An open model for hypermedia with dynamic linking. In *Hypertext: The Proceedings of The European Conference on Hypertext Concepts, Systems and Applications.*, INRIA, France, 1990.

[3] D. W. Huang, Y. Xu, A. Trotman, and S. Geva. Overview of inex 2007 link the wiki track. In N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, editors, *Focused Access to XML Documents*, pages 373–387. Springer-Verlag, Berlin, Heidelberg, 2008. ISBN 978-3-540-85901-7. doi: 10.1007/978-3-540-85902-4_32.

[4] S. McDonald and R. J. Stevenson. Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers*, 10(2):129 – 142, 1998. ISSN 0953-5438. doi: 10.1016/S0953-5438(98)00017-4.

[5] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458150.

[6] M. Otter and H. Johnson. Lost in hyperspace: metrics and mental models. *Interacting with Computers*, 13(1):1 – 40, 2000. ISSN 0953-5438. doi: 10.1016/S0953-5438(00)00030-8.

[7] P. A. Smith. Towards a practical measure of hypertext usability. *Interacting with Computers*, 8(4):365 – 381, 1996. ISSN 0953-5438. doi: 10.1016/S0953-5438(97)83779-4.

[8] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 315–323, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291017.

[9] C. Y. Wei, M. B. Evans, M. Eliot, J. Barrick, B. Maust, and J. H. Spyridakis. Influencing web-browsing behavior with intriguing and informative hyperlink wording. *Journal of Information Science*, 31(5):433–445, 2005. doi: 10.1177/0165551505055703.